COMPSCI 514: Algorithms for Data Science

Cameron Musco

University of Massachusetts Amherst. Spring 2026.
Lecture 12

Logistics

- Quiz this week due Monday at 8pm.
- Several survey style questions. Due to Canvas limitations, some options may be marked incorrect but still will receive full credit.
- So just ignore the grading on the survey style questions.

Summary

Last Class:

- Intro to dimensionality reduction.
- Low-distortion embeddings and the Johnson-Lindenstrauss (JL) Lemma.
- Most of the proof of the JL lemma.

This Class:

- Finish JL Lemma proof.
- Example application to clustering.
- Start on data dependent dimensionality reduction.
  - PCA

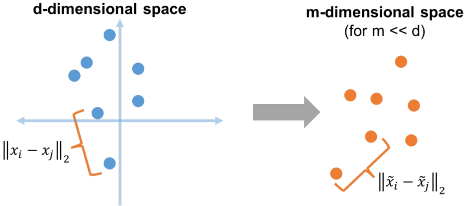
The Johnson-Lindenstrauss Lemma

Johnson-Lindenstrauss Lemma: For any set of points $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$ and $\epsilon > 0$ there exists a linear map $\mathbf{\Pi} : \mathbb{R}^d \to \mathbb{R}^m$ such that $m = O\left(\frac{\log n}{\epsilon^2}\right)$ and letting $\tilde{x}_i = \mathbf{\Pi}\vec{x}_i$:

$m << d$

$$\text{For all } i,j: \ (1-\epsilon)\|\vec{x}_i - \vec{x}_j\|_2 \leq \underline{\|\tilde{x}_i - \tilde{x}_j\|_2 \leq (1+\epsilon)}\|\vec{x}_i - \vec{x}_j\|_2.$$

Further, if $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$ has each entry chosen i.i.d. from $\mathcal{N}(0, 1/m)$, it satisfies the guarantee with high probability.



Distributional JL

The Johnson-Lindenstrauss Lemma is a direct consequence of a closely related lemma:

Distributional JL Lemma: Let $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$ have each entry chosen i.i.d. as $\mathcal{N}(0, 1/m)$. If we set $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, then for any $\vec{y} \in \mathbb{R}^d$, with probability $\geq 1 - \delta$

$$(1 - \underline{\epsilon)\|\vec{y}\|}_2 \leq \underline{\|\mathbf{\Pi}\vec{y}\|}_2 \leq (1 + \epsilon)\underline{\|\vec{y}\|}_2$$

Applying a random matrix $\mathbf{\Pi}$ to any vector $\vec{y}$ preserves $\vec{y}$'s norm with high probability.

Distributional JL Lemma $\implies$ JL Lemma: Apply Distributional-JL to $y_{ij} = x_i - x_j$ for all $\binom{n}{2}$ pairs x_i, x_j. Then take a union bound to say distance is preserved between all pairs with high probability.

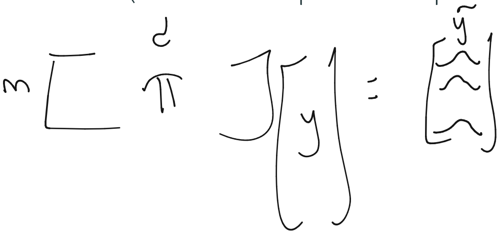
Distributional JL Proof

Letting $\mathbf{\Pi} \in \mathbb{R}^{d \times m}$ have each entry chosen i.i.d. as $\mathcal{N}(0, 1/m)$, for any $\vec{y} \in \mathbb{R}^d$, letting $\tilde{\mathbf{y}} = \mathbf{\Pi}\vec{y}$:

$$\underline{\tilde{\mathbf{y}}(j)} \sim \mathcal{N}(0, \|\vec{y}\|_2^2/m) \text{ and } \mathbb{E}[\|\tilde{\mathbf{y}}\|_2^2] = \|\vec{y}\|_2^2$$

$$\|\tilde{y}\|_2^2 \approx \|y\|_2^2$$

$\|\tilde{\mathbf{y}}\|_2^2 = \sum_{i=1}^m \tilde{\mathbf{y}}(j)^2$ a Chi-Squared random variable with m degrees of freedom (a sum of m squared independent Gaussians).



$\vec{y} \in \mathbb{R}^d$: arbitrary vector, $\tilde{\mathbf{y}} \in \mathbb{R}^m$: compressed vector, $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$: random projection mapping $\vec{y} \rightarrow \tilde{\mathbf{y}}$. d: original dimension. m: compressed dimension, ϵ: embedding error, δ: embedding failure prob.

Distributional JL Proof

Letting $\mathbf{\Pi} \in \mathbb{R}^{d \times m}$ have each entry chosen i.i.d. as $\mathcal{N}(0, 1/m)$, for any $\vec{y} \in \mathbb{R}^d$, letting $\tilde{\mathbf{y}} = \mathbf{\Pi}\vec{y}$:

$$\tilde{\mathbf{y}}(j) \sim \mathcal{N}(0, \|\vec{y}\|_2^2/m) \text{ and } \mathbb{E}[\|\tilde{\mathbf{y}}\|_2^2] = \|\vec{y}\|_2^2$$

$\|\tilde{\mathbf{y}}\|_2^2 = \sum_{i=1}^m \tilde{\mathbf{y}}(j)^2$ a Chi-Squared random variable with m degrees of freedom (a sum of m squared independent Gaussians).

Lemma: (Chi-Squared Concentration) Letting $\mathbf{Z}$ be a Chi-Squared random variable with m degrees of freedom,

$$\Pr\left[|\mathbf{Z} - \mathbb{E}\mathbf{Z}| \geq \epsilon \mathbb{E}\mathbf{Z}\right] \leq 2e^{-m\epsilon^2/8}.$$

$$\Pr\left[\left|\|\tilde{y}\|_2^2 - \|y\|_2^2\right| \geq \epsilon\|y\|_2^2\right] \leq 2e^{-m\epsilon^2/8} = 2e^{-O(\log(1/\delta))} \leq \delta$$

$$(1-\epsilon)\|y\|_2^2 \leq \|\tilde{y}\|_2^2 \leq (1+\epsilon)\|y\|_2^2$$

$$m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$$

$\vec{y} \in \mathbb{R}^d$: arbitrary vector, $\tilde{\mathbf{y}} \in \mathbb{R}^m$: compressed vector, $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$: random projection mapping $\vec{y} \to \tilde{\mathbf{y}}$. d: original dimension. m: compressed dimension, ϵ: embedding error, δ: embedding failure prob.

Distributional JL Proof

Letting $\boldsymbol{\Pi} \in \mathbb{R}^{d \times m}$ have each entry chosen i.i.d. as $\mathcal{N}(0, 1/m)$, for any $\vec{y} \in \mathbb{R}^d$, letting $\tilde{\mathbf{y}} = \boldsymbol{\Pi}\vec{y}$:

$$\tilde{\mathbf{y}}(j) \sim \mathcal{N}(0, \|\vec{y}\|_2^2/m) \text{ and } \mathbb{E}[\|\tilde{\mathbf{y}}\|_2^2] = \|\vec{y}\|_2^2$$

$\|\tilde{\mathbf{y}}\|_2^2 = \sum_{i=1}^m \tilde{\mathbf{y}}(j)^2$ a Chi-Squared random variable with m degrees of freedom (a sum of m squared independent Gaussians).

Lemma: (Chi-Squared Concentration) Letting $\mathbf{Z}$ be a Chi-Squared random variable with m degrees of freedom,

$$\Pr\left[|\mathbf{Z} - \mathbb{E}\mathbf{Z}| \geq \epsilon \mathbb{E}\mathbf{Z}\right] \leq 2e^{-m\epsilon^2/8}.$$

If we set $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, with probability $1 - O(e^{-\log(1/\delta)}) \geq 1 - \delta$:

$$(1-\epsilon)\|\vec{y}\|_2^2 \leq \|\tilde{\mathbf{y}}\|_2^2 \leq (1+\epsilon)\|\vec{y}\|_2^2.$$

$\vec{y} \in \mathbb{R}^d$: arbitrary vector, $\tilde{\mathbf{y}} \in \mathbb{R}^m$: compressed vector, $\boldsymbol{\Pi} \in \mathbb{R}^{m \times d}$: random projection mapping $\vec{y} \to \tilde{\mathbf{y}}$. d: original dimension. m: compressed dimension, ϵ: embedding error, δ: embedding failure prob.

Distributional JL Proof

Letting $\mathbf{\Pi} \in \mathbb{R}^{d \times m}$ have each entry chosen i.i.d. as $\mathcal{N}(0, 1/m)$, for any $\vec{y} \in \mathbb{R}^d$, letting $\tilde{\mathbf{y}} = \mathbf{\Pi}\vec{y}$:

$$\tilde{\mathbf{y}}(j) \sim \mathcal{N}(0, \|\vec{y}\|_2^2/m) \text{ and } \mathbb{E}[\|\tilde{\mathbf{y}}\|_2^2] = \|\vec{y}\|_2^2$$

$\|\tilde{\mathbf{y}}\|_2^2 = \sum_{i=1}^m \tilde{\mathbf{y}}(j)^2$ a Chi-Squared random variable with m degrees of freedom (a sum of m squared independent Gaussians).

Lemma: (Chi-Squared Concentration) Letting $\mathbf{Z}$ be a Chi-Squared random variable with m degrees of freedom,

$$\Pr\left[|\mathbf{Z} - \mathbb{E}\mathbf{Z}| \geq \epsilon \mathbb{E}\mathbf{Z}\right] \leq 2e^{-m\epsilon^2/8}.$$

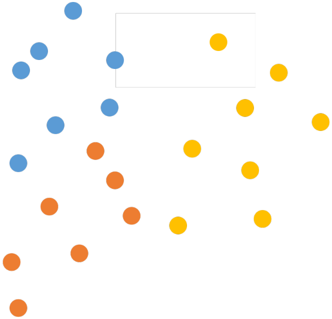
If we set $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, with probability $1 - O(e^{-\log(1/\delta)}) \geq 1 - \delta$:

$$(1-\epsilon)\|\vec{y}\|_2^2 \leq \|\tilde{\mathbf{y}}\|_2^2 \leq (1+\epsilon)\|\vec{y}\|_2^2.$$

Gives the distributional JL Lemma and thus the classic JL Lemma.

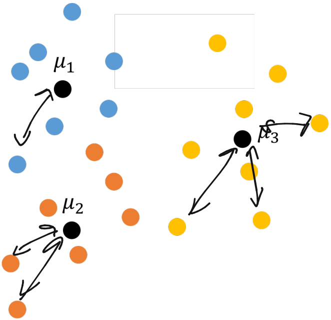
Example Application: *k*-means clustering

Goal: Separate n points in d dimensional space into k groups.



Example Application: k-means clustering

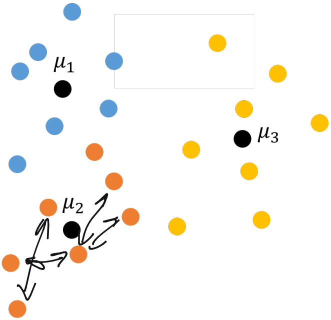
Goal: Separate n points in d dimensional space into k groups.



k-means Objective: $Cost(\mathcal{C}_1, \ldots, \mathcal{C}_k) = \min_{\mathcal{C}_1,\ldots\mathcal{C}_k} \sum_{j=1}^{k} \sum_{\vec{x} \in \mathcal{C}_k} \|\vec{x} - \mu_j\|_2^2.$

Example Application: k-means clustering

Goal: Separate n points in d dimensional space into k groups.



k-means Objective: $Cost(\mathcal{C}_1, \ldots, \mathcal{C}_k) = \min_{\mathcal{C}_1, \ldots \mathcal{C}_k} \sum_{j=1}^{k} \sum_{\vec{x} \in \mathcal{C}_k} \|\vec{x} - \mu_j\|_2^2.$

Write in terms of distances:

$$Cost(\mathcal{C}_1, \ldots, \mathcal{C}_k) = \min_{\mathcal{C}_1, \ldots \mathcal{C}_k} \sum_{j=1}^{k} \sum_{\vec{x}_1, \vec{x}_2 \in \mathcal{C}_k} \|\vec{x}_1 - \vec{x}_2\|_2^2$$

Example Application: *k*-means clustering

k-means Objective: $Cost(\mathcal{C}_1, \ldots, \mathcal{C}_k) = \min_{\mathcal{C}_1, \ldots \mathcal{C}_k} \sum_{j=1}^{k} \sum_{\vec{x}_1, \vec{x}_2 \in \mathcal{C}_k} \|\vec{x}_1 - \vec{x}_2\|_2^2$

Example Application: k-means clustering

k-means Objective: $Cost(\mathcal{C}_1, \ldots, \mathcal{C}_k) = \min_{\mathcal{C}_1, \ldots \mathcal{C}_k} \sum_{j=1}^{k} \sum_{\vec{x}_1, \vec{x}_2 \in \mathcal{C}_k} \|\vec{x}_1 - \vec{x}_2\|_2^2$

If we randomly project to $\underline{m = O\left(\frac{\log n}{\epsilon^2}\right)}$ dimensions, for all pairs $\vec{x}_1, \vec{x}_2$,

$$\underline{(1-\epsilon)\|\vec{x}_1 - \vec{x}_2\|_2^2 \leq \|\tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2\|_2^2 \leq (1+\epsilon)\|\vec{x}_1 - \vec{x}_2\|_2^2}$$

Example Application: k-means clustering

k-means Objective: $Cost(\mathcal{C}_1,\ldots,\mathcal{C}_k) = \min_{\mathcal{C}_1,\ldots\mathcal{C}_k} \sum_{j=1}^{k} \sum_{\vec{x}_1,\vec{x}_2\in\mathcal{C}_k} \|\vec{x}_1-\vec{x}_2\|_2^2$

If we randomly project to $m = O\left(\frac{\log n}{\epsilon^2}\right)$ dimensions, for all pairs $\vec{x}_1,\vec{x}_2$,

$$(1-\epsilon)\|\vec{x}_1-\vec{x}_2\|_2^2 \le \|\tilde{\mathbf{x}}_1-\tilde{\mathbf{x}}_2\|_2^2 \le (1+\epsilon)\|\vec{x}_1-\vec{x}_2\|_2^2 \implies$$

Letting $\overline{Cost}(\mathcal{C}_1,\ldots,\mathcal{C}_k) = \min_{\mathcal{C}_1,\ldots\mathcal{C}_k} \sum_{j=1}^{k} \sum_{\tilde{\mathbf{x}}_1,\tilde{\mathbf{x}}_2\in\mathcal{C}_k} \|\tilde{\mathbf{x}}_1-\tilde{\mathbf{x}}_2\|_2^2$

$$(1-\epsilon)Cost(\mathcal{C}_1,\ldots,\mathcal{C}_k) \le \overline{Cost}(\mathcal{C}_1,\ldots,\mathcal{C}_k) \le (1+\epsilon)Cost(\mathcal{C}_1,\ldots,\mathcal{C}_k).$$

Example Application: k-means clustering

k-means Objective: $Cost(\mathcal{C}_1, \ldots, \mathcal{C}_k) = \min_{\mathcal{C}_1,\ldots\mathcal{C}_k} \sum_{j=1}^{k} \sum_{\vec{x}_1, \vec{x}_2 \in \mathcal{C}_k} \|\vec{x}_1 - \vec{x}_2\|_2^2$

If we randomly project to $m = O\left(\frac{\log n}{\epsilon^2}\right)$ dimensions, for all pairs $\vec{x}_1, \vec{x}_2$,

$$(1-\epsilon)\|\vec{x}_1 - \vec{x}_2\|_2^2 \leq \|\tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2\|_2^2 \leq (1+\epsilon)\|\vec{x}_1 - \vec{x}_2\|_2^2 \implies$$

Letting $\overline{Cost}(\mathcal{C}_1, \ldots, \mathcal{C}_k) = \min_{\mathcal{C}_1,\ldots\mathcal{C}_k} \sum_{j=1}^{k} \sum_{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2 \in \mathcal{C}_k} \|\tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2\|_2^2$

$$(1-\epsilon)Cost(\mathcal{C}_1, \ldots, \mathcal{C}_k) \leq \overline{Cost}(\mathcal{C}_1, \ldots, \mathcal{C}_k) \leq (1+\epsilon)Cost(\mathcal{C}_1, \ldots, \mathcal{C}_k).$$

Upshot: Can cluster in m dimensional space (much more efficiently) and minimize $\overline{Cost}(\mathcal{C}_1, \ldots, \mathcal{C}_k)$. The optimal set of clusters will have true cost within $\frac{1+\epsilon}{1-\epsilon} \approx 1 + 2\epsilon$ times the true optimal. Good exercise to prove this.

we prove

Dist. JL : random Π preserves lengths of vectors whp

↓

JL : " preserves distances b/t vectors whp

Data Dependent Dimensionality Reduction

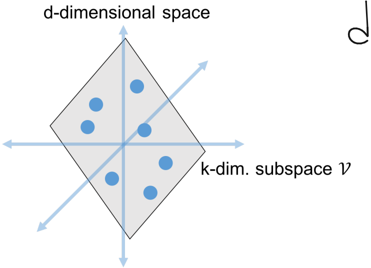
Data Dependent Dimensionality Reduction

- The JL Lemma is very general – it applies to any input data set.
- The compression Π is linear and data-oblivious. This is very useful for computational efficiency, and leads to applications in streaming and distributed settings.
- However, often we can exploit properties about the underlying input data to get more accurate compressions than what JL gives.
- Leads to a variety of data dependent dimensionality reduction methods.

-PCA and its many varietys

Embedding with Assumptions

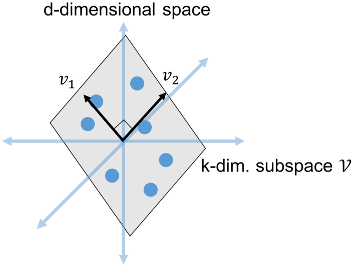
Assume that data points $\vec{x}_1, \ldots, \vec{x}_n$ lie in any k-dimensional subspace $\mathcal{V}$ of $\mathbb{R}^d$.



$d \to k$

Embedding with Assumptions

Assume that data points $\vec{x}_1, \ldots, \vec{x}_n$ lie in any k-dimensional subspace $\mathcal{V}$ of $\mathbb{R}^d$.



$$\begin{array}{cc} k & \\ \begin{bmatrix} v_1 \cdots v_k \\ V \end{bmatrix} & d \end{array}$$

Claim: Let $\vec{v}_1, \ldots, \vec{v}_k$ be an $\underline{\text{orthonormal}}$ basis for $\mathcal{V}$ and $\mathbf{V} \in \mathbb{R}^{d \times k}$ be the matrix with these vectors as its columns. For all $\vec{x}_i, \vec{x}_j$:

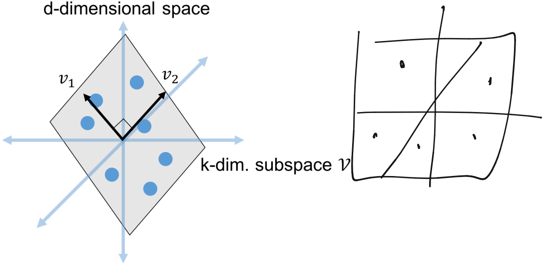
$$\|\mathbf{V}^T\vec{x}_i - \mathbf{V}^T\vec{x}_j\|_2 = \|\vec{x}_i - \vec{x}_j\|_2.$$

$$\begin{array}{c} d \\ k\begin{bmatrix} V^T \end{bmatrix} \end{array} \begin{bmatrix} x_i \end{bmatrix} - \begin{bmatrix} V^T \end{bmatrix}\begin{bmatrix} x_j \end{bmatrix}$$

$$k \begin{bmatrix} V^T x_i \end{bmatrix} - \begin{bmatrix} V^T x_j \end{bmatrix}$$

Embedding with Assumptions

Assume that data points $\vec{x}_1, \ldots, \vec{x}_n$ lie in any k-dimensional subspace $\mathcal{V}$ of $\mathbb{R}^d$.



Claim: Let $\vec{v}_1, \ldots, \vec{v}_k$ be an orthonormal basis for $\mathcal{V}$ and $\mathbf{V} \in \mathbb{R}^{d \times k}$ be the matrix with these vectors as its columns. For all $\vec{x}_i, \vec{x}_j$:

$$\|\mathbf{V}^T\vec{x}_i - \mathbf{V}^T\vec{x}_j\|_2 = \|\vec{x}_i - \vec{x}_j\|_2.$$

- $\mathbf{V}^T \in \mathbb{R}^{k \times d}$ is a linear embedding of $\vec{x}_1, \ldots, \vec{x}_n$ into k dimensions with no distortion.

Very Simple Example

$d = 3$ $\quad \mathbb{R}^3$ $\quad k = 2$ $\quad$ subspace x, y-plane

$$\underset{V_1}{\begin{bmatrix}1\\0\\0\end{bmatrix}} \quad \underset{V_2}{\begin{bmatrix}0\\1\\0\end{bmatrix}}$$

orthonormal basis for x, y plane

$$V = \begin{bmatrix}1 & 0\\0 & 1\\0 & 0\end{bmatrix}$$

$$\overset{V^T}{\begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\end{bmatrix}} \begin{bmatrix}x(1)\\x(2)\\x(3)\end{bmatrix} = \begin{bmatrix}x(1)\\x(2)\end{bmatrix}$$

$$V^T x = \begin{bmatrix}x(1)\\x(2)\end{bmatrix}$$

if x_i, x_j are in xy-plane

$$\|V^T x_i - V^T x_j\|_2 = \|x_i - x_j\|$$

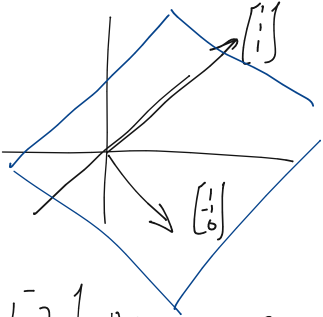
$$\begin{bmatrix}1\\2\\0\end{bmatrix} - \begin{bmatrix}3\\4\\0\end{bmatrix}$$

Another Example

$d = 3$ $\qquad$ $k = 2$

$$\begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{2} \\ 1/\sqrt{3} & -1/\sqrt{2} \\ 1/\sqrt{3} & 0 \end{bmatrix}$$

$$\begin{matrix}1\\1\\1\end{matrix} + \begin{matrix}1\\-1\\0\end{matrix} = \begin{bmatrix}2\\0\\1\end{bmatrix}$$



$$\left\| \begin{bmatrix}2\\0\\1\end{bmatrix} - \begin{bmatrix}2\\2\\2\end{bmatrix} \right\|_2^2 = 0^2 + 2^2 + (-1)^2 = 5$$

$$\begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \end{bmatrix} \begin{bmatrix}2\\0\\1\end{bmatrix} = \begin{bmatrix}\sqrt{3}\\ \sqrt{2}\end{bmatrix}$$

$$\begin{bmatrix} \cdots \end{bmatrix} \begin{bmatrix}2\\2\\2\end{bmatrix} = \begin{bmatrix}2\sqrt{3}\\ 0\end{bmatrix}$$

$$\left\| \begin{bmatrix}\sqrt{3}\\ \sqrt{2}\end{bmatrix} - \begin{bmatrix}2\sqrt{3}\\ 0\end{bmatrix} \right\|_2^2 = (\sqrt{3})^2 + (\sqrt{2})^2 = 5$$

Dot Product Transformation

Claim: Let $\vec{v}_1, \ldots, \vec{v}_k$ be an orthonormal basis for $\mathcal{V}$ and $\mathbf{V} \in \mathbb{R}^{d \times k}$ be the matrix with these vectors as its columns. For all $\vec{x}_i, \vec{x}_j \in \mathcal{V}$:

$$\|\mathbf{V}^T\vec{x}_i - \mathbf{V}^T\vec{x}_j\|_2 = \|\vec{x}_i - \vec{x}_j\|_2.$$

$$\|c_i - c_j\|_2 = \|x_i - x_j\|_2$$

$$\|V^T V c_i - V^T V c_j\|_2 = \|c_i - c_j\|_2$$

$$X \in V$$

$$X = c_1 \cdot \vec{v}_1 + c_2 \vec{v}_2 \cdots + c_k \vec{v}_k$$

$$\begin{array}{cc} & 1 \\ d & \begin{bmatrix} X \end{bmatrix} \end{array} = \begin{array}{cc} & k \\ d & \begin{bmatrix} v_1 \ldots V \ldots v_k \end{bmatrix} \end{array} \begin{array}{cc} 1 & \\ \begin{bmatrix} \vec{c} \end{bmatrix} & k \end{array}$$

$$\exists c \in \mathbb{R}^k \quad X = Vc$$

$$\begin{bmatrix} V^T V \end{bmatrix}$$

$$\begin{array}{cc} & d \\ k & \begin{bmatrix} V^T \end{bmatrix} \end{array} \begin{array}{cc} k & \\ \begin{bmatrix} V \end{bmatrix} & d \end{array} = \begin{array}{cc} k & \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} & k \end{array}$$

$$V^T V$$

$$(V^T V)_{ij} = v_i^T v_j = 1 \text{ if } i = j$$

$$0 \text{ o.w. by orthonormality}$$

Dot Product Transformation

Claim: Let $\vec{v}_1, \ldots, \vec{v}_k$ be an orthonormal basis for $\mathcal{V}$ and $\mathbf{V} \in \mathbb{R}^{d \times k}$ be the matrix with these vectors as its columns. For all $\vec{x}_i, \vec{x}_j \in \mathcal{V}$:

$$\|\mathbf{V}^T\vec{x}_i - \mathbf{V}^T\vec{x}_j\|_2 = \|\vec{x}_i - \vec{x}_j\|_2.$$

$x_i = Vc_i \quad x_j = Vc_j$

$$\|y\|_2^2 = \sum y(i)^2$$

$$= \langle y, y \rangle$$

$$= y^T y$$

$$\|c_i - c_j\|_2^2 = \|x_i - x_j\|_2^2$$

$$= \|Vc_i - Vc_j\|_2^2$$

$$= (Vc_i - Vc_j)^T (Vc_i - Vc_j)$$

$$= [V(c_i - c_j)]^T V(c_i - c_j)$$

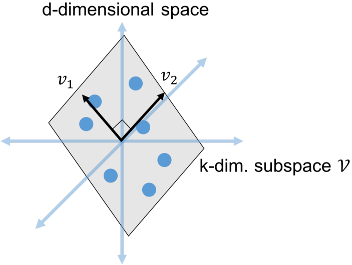
$$(c_i - c_j)^T \overset{I}{\cancel{V^T V}} (c_i - c_j)$$

$$(c_i - c_j)^T (c_i - c_j)$$

$$= \|c_i - c_j\|_2^2$$

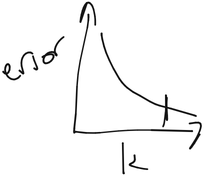
Embedding with Assumptions

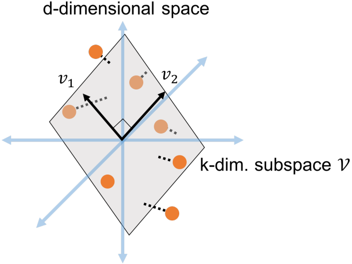
Main Focus of Upcoming Classes: Assume that data points $\vec{x}_1, \ldots, \vec{x}_n$ lie close to any k-dimensional subspace $\mathcal{V}$ of $\mathbb{R}^d$.



Embedding with Assumptions

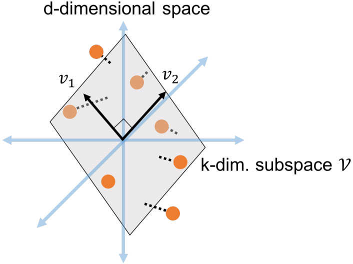
Main Focus of Upcoming Classes: Assume that data points $\vec{x}_1, \ldots, \vec{x}_n$ lie close to any k-dimensional subspace $\mathcal{V}$ of $\mathbb{R}^d$.





Embedding with Assumptions

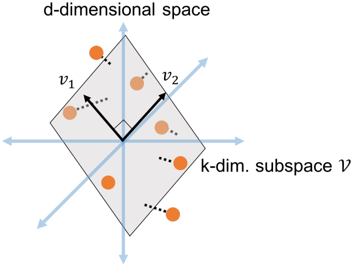
Main Focus of Upcoming Classes: Assume that data points $\vec{x}_1, \ldots, \vec{x}_n$ lie close to any k-dimensional subspace $\mathcal{V}$ of $\mathbb{R}^d$.



Letting $\vec{v}_1, \ldots, \vec{v}_k$ be an orthonormal basis for $\mathcal{V}$ and $\mathbf{V} \in \mathbb{R}^{d \times k}$ be the matrix with these vectors as its columns, $\mathbf{V}^T \vec{x}_i \in \mathbb{R}^k$ is still a good embedding for $x_i \in \mathbb{R}^d$.

Embedding with Assumptions

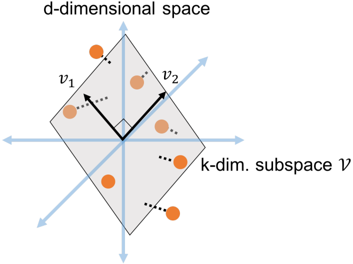
Main Focus of Upcoming Classes: Assume that data points $\vec{x}_1, \ldots, \vec{x}_n$ lie close to any k-dimensional subspace $\mathcal{V}$ of $\mathbb{R}^d$.



Letting $\vec{v}_1, \ldots, \vec{v}_k$ be an orthonormal basis for $\mathcal{V}$ and $\mathbf{V} \in \mathbb{R}^{d \times k}$ be the matrix with these vectors as its columns, $\mathbf{V}^T \vec{x}_i \in \mathbb{R}^k$ is still a good embedding for $x_i \in \mathbb{R}^d$. The key idea behind low-rank approximation and principal component analysis (PCA).

Embedding with Assumptions

Main Focus of Upcoming Classes: Assume that data points $\vec{x}_1, \ldots, \vec{x}_n$ lie close to any k-dimensional subspace $\mathcal{V}$ of $\mathbb{R}^d$.



Letting $\vec{v}_1, \ldots, \vec{v}_k$ be an orthonormal basis for $\mathcal{V}$ and $\mathbf{V} \in \mathbb{R}^{d \times k}$ be the matrix with these vectors as its columns, $\mathbf{V}^T\vec{x}_i \in \mathbb{R}^k$ is still a good embedding for $x_i \in \mathbb{R}^d$. The key idea behind low-rank approximation and principal component analysis (PCA).

- How do we find $\mathcal{V}$ and $\mathbf{V}$?
- How good is the embedding?

Low-Rank Factorization

Claim: $\vec{x}_1, \ldots, \vec{x}_n$ lie in a k-dimensional subspace $\mathcal{V} \Leftrightarrow$ the data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ has rank $\leq k$.

$$\begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \quad X \\ x_n^T \end{bmatrix}$$

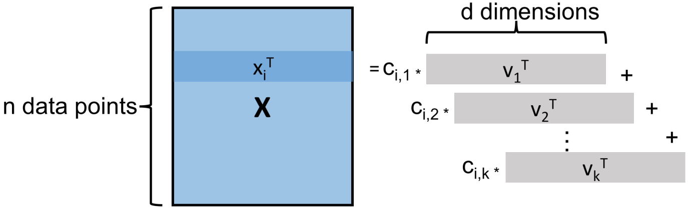
$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Low-Rank Factorization

Claim: $\vec{x}_1, \ldots, \vec{x}_n$ lie in a k-dimensional subspace $\mathcal{V} \Leftrightarrow$ the data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ has rank $\leq k$.

- Letting $\vec{v}_1, \ldots, \vec{v}_k$ be an orthonormal basis for $\mathcal{V}$, can write $\vec{x}_i$ as:

$$\vec{x}_i = \mathbf{V}\vec{c}_i = c_{i,1} \cdot \vec{v}_1 + c_{i,2} \cdot \vec{v}_2 + \ldots + c_{i,k} \cdot \vec{v}_k.$$



$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

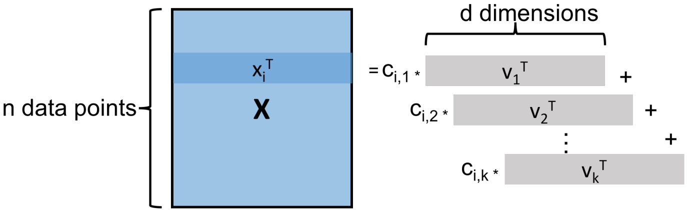
Low-Rank Factorization

Claim: $\vec{x}_1, \ldots, \vec{x}_n$ lie in a k-dimensional subspace $\mathcal{V} \Leftrightarrow$ the data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ has rank $\leq k$.

- Letting $\vec{v}_1, \ldots, \vec{v}_k$ be an orthonormal basis for $\mathcal{V}$, can write $\vec{x}_i$ as:

$$\vec{x}_i = \mathbf{V}\vec{c}_i = c_{i,1} \cdot \vec{v}_1 + c_{i,2} \cdot \vec{v}_2 + \ldots + c_{i,k} \cdot \vec{v}_k.$$

- So $\vec{v}_1, \ldots, \vec{v}_k$ span the rows of $\mathbf{X}$ and thus $\mathbf{rank}(\mathbf{X}) \leq k$.



$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

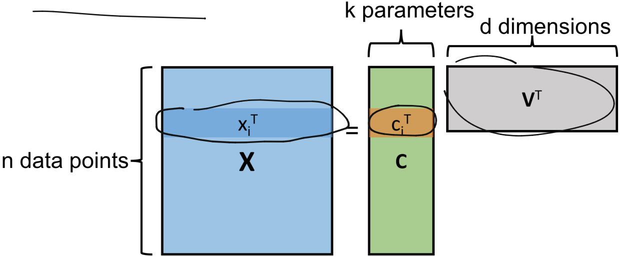
Claim: $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$ lie in a k-dimensional subspace $\mathcal{V} \Leftrightarrow$ the data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ has rank $\leq k$.

- Every data point $\vec{x}_i$ (row of $\mathbf{X}$) can be written as $\vec{x}_i = \mathbf{V}\vec{c}_i = c_{i,1} \cdot \vec{v}_1 + \ldots + c_{i,k} \cdot \vec{v}_k$.

$\vec{x}_1, \ldots, \vec{x}_n$: data points (in $\mathbb{R}^d$), $\mathcal{V}$: k-dimensional subspace of $\mathbb{R}^d$, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Claim: $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$ lie in a k-dimensional subspace $\mathcal{V} \Leftrightarrow$ the data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ has rank $\leq k$.

- Every data point $\vec{x}_i$ (row of $\mathbf{X}$) can be written as $\vec{x}_i = \mathbf{V}\vec{c}_i = \underline{c_{i,1} \cdot \vec{v}_1 + \ldots + c_{i,k}} \cdot \vec{v}_k$.



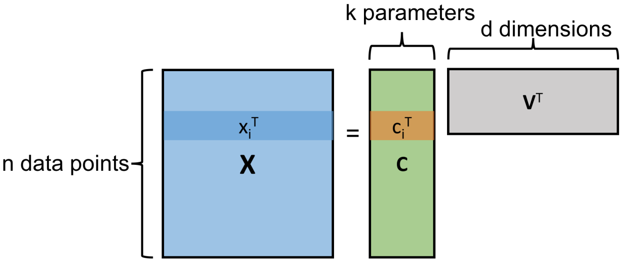
$x_i = Vc_i$

$x_i^T = c_i^T V^T$

$\vec{x}_1, \ldots, \vec{x}_n$: data points (in $\mathbb{R}^d$), $\mathcal{V}$: k-dimensional subspace of $\mathbb{R}^d$, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Claim: $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$ lie in a k-dimensional subspace $\mathcal{V} \Leftrightarrow$ the data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ has rank $\leq k$.

- Every data point $\vec{x}_i$ (row of $\mathbf{X}$) can be written as $\vec{x}_i = \mathbf{V}\vec{c}_i = c_{i,1} \cdot \vec{v}_1 + \ldots + c_{i,k} \cdot \vec{v}_k$.

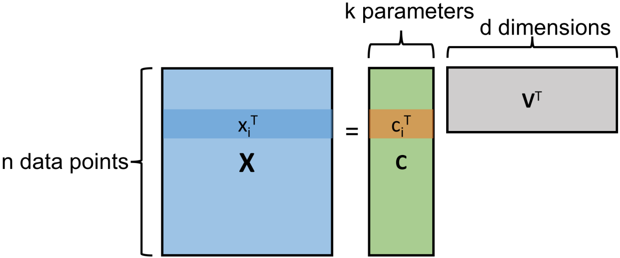


- $\mathbf{X}$ can be represented by $(n+d) \cdot k$ parameters vs. $n \cdot d$.

$\vec{x}_1, \ldots, \vec{x}_n$: data points (in $\mathbb{R}^d$), $\mathcal{V}$: k-dimensional subspace of $\mathbb{R}^d$, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Claim: $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$ lie in a k-dimensional subspace $\mathcal{V} \Leftrightarrow$ the data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ has rank $\leq k$.

- Every data point $\vec{x}_i$ (row of $\mathbf{X}$) can be written as $\vec{x}_i = \mathbf{V}\vec{c}_i = c_{i,1} \cdot \vec{v}_1 + \ldots + c_{i,k} \cdot \vec{v}_k$.

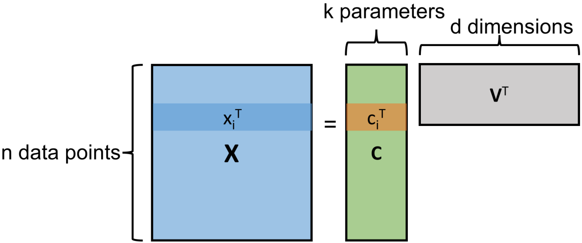


- $\mathbf{X}$ can be represented by $(n+d) \cdot k$ parameters vs. $n \cdot d$.
- The rows of $\mathbf{X}$ are spanned by k vectors: the columns of $\mathbf{V} \implies$ the columns of $\mathbf{X}$ are spanned by k vectors: the columns of $\mathbf{C}$.

$\vec{x}_1, \ldots, \vec{x}_n$: data points (in $\mathbb{R}^d$), $\mathcal{V}$: k-dimensional subspace of $\mathbb{R}^d$, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Low-Rank Factorization

Claim: If $\vec{x}_1, \ldots, \vec{x}_n$ lie in a k-dimensional subspace with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be written as $\mathbf{X} = \mathbf{C}\mathbf{V}^T$.

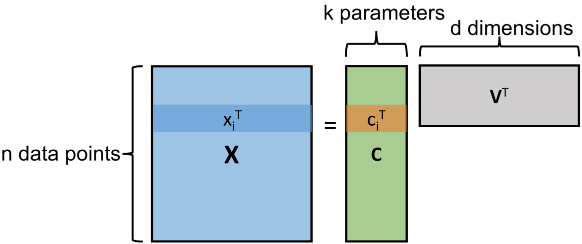


$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Low-Rank Factorization

Claim: If $\vec{x}_1, \ldots, \vec{x}_n$ lie in a k-dimensional subspace with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be written as $\mathbf{X} = \mathbf{C}\mathbf{V}^T$.

X, V



Exercise: What is this coefficient matrix $\mathbf{C}$? **Hint:** Use that $\mathbf{V}^T\mathbf{V} = \mathbf{I}$.

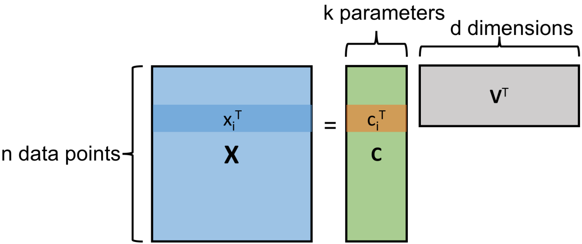
C

$$X = CV^T$$
$$XV = CV^TV \Rightarrow \boxed{XV = C}$$

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Low-Rank Factorization

Claim: If $\vec{x}_1, \ldots, \vec{x}_n$ lie in a k-dimensional subspace with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be written as $\mathbf{X} = \mathbf{C}\mathbf{V}^T$.



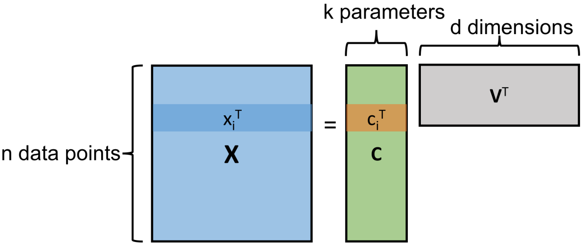
Exercise: What is this coefficient matrix C? **Hint:** Use that $\mathbf{V}^T\mathbf{V} = \mathbf{I}$.

- $\mathbf{X} = \mathbf{C}\mathbf{V}^T \implies \mathbf{X}\mathbf{V} = \mathbf{C}\mathbf{V}^T\mathbf{V}$

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Low-Rank Factorization

Claim: If $\vec{x}_1, \ldots, \vec{x}_n$ lie in a k-dimensional subspace with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be written as $\mathbf{X} = \mathbf{C}\mathbf{V}^T$.



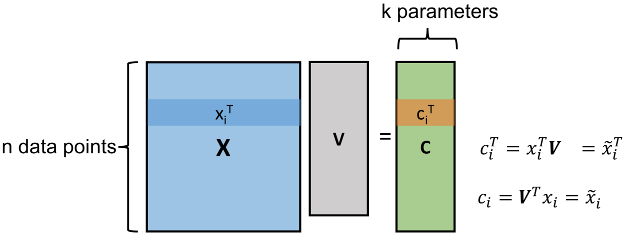
Exercise: What is this coefficient matrix C? **Hint:** Use that $\mathbf{V}^T\mathbf{V} = \mathbf{I}$.

- $\mathbf{X} = \mathbf{C}\mathbf{V}^T \implies \mathbf{X}\mathbf{V} = \mathbf{C}\mathbf{V}^T\mathbf{V} \implies \mathbf{X}\mathbf{V} = \mathbf{C}$

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Low-Rank Factorization

Claim: If $\vec{x}_1, \ldots, \vec{x}_n$ lie in a k-dimensional subspace with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be written as $\mathbf{X} = \mathbf{C}\mathbf{V}^T$.



Exercise: What is this coefficient matrix $\mathbf{C}$? **Hint:** Use that $\mathbf{V}^T\mathbf{V} = \mathbf{I}$.

- $\mathbf{X} = \mathbf{C}\mathbf{V}^T \implies \mathbf{X}\mathbf{V} = \mathbf{C}\mathbf{V}^T\mathbf{V} \implies \mathbf{X}\mathbf{V} = \mathbf{C}$

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Projection View

Claim: If $\vec{x}_1, \ldots, \vec{x}_n$ lie in a k-dimensional subspace $\mathcal{V}$ with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be written as

$$\underset{\substack{n\times d \\ \text{original data}}}{\mathbf{X}} = \underset{\substack{C = XV \text{ is } n\times k \\ \text{compressed data}}}{\mathbf{C}}\mathbf{V}^T.$$

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Projection View

Claim: If $\vec{x}_1, \ldots, \vec{x}_n$ lie in a k-dimensional subspace $\mathcal{V}$ with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be written as

$$\mathbf{X} = \mathbf{X}\mathbf{V}\mathbf{V}^T.$$

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Projection View

Claim: If $\vec{x}_1, \ldots, \vec{x}_n$ lie in a k-dimensional subspace $\mathcal{V}$ with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be written as

$$\mathbf{X} = \mathbf{X}\mathbf{V}\mathbf{V}^T.$$

$$k\underset{k}{\left[\; V^T V \;\right]} \qquad VV^T = d\overset{d}{\left[\qquad\quad\right]}$$

- $\mathbf{V}\mathbf{V}^T$ is a projection matrix, which projects vectors onto the subspace $\mathcal{V}$.

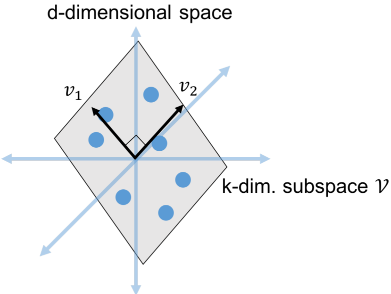
$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Projection View

Claim: If $\vec{x}_1, \ldots, \vec{x}_n$ lie in a k-dimensional subspace $\mathcal{V}$ with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be written as

$$\mathbf{X} = \mathbf{X}\mathbf{V}\mathbf{V}^T.$$

- $\mathbf{V}\mathbf{V}^T$ is a projection matrix, which projects vectors onto the subspace $\mathcal{V}$.



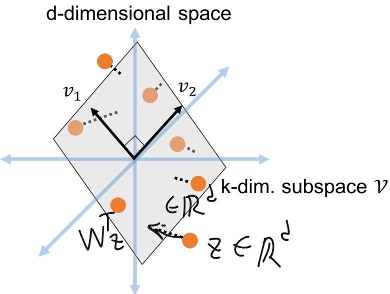
$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Projection View

Claim: If $\vec{x}_1, \ldots, \vec{x}_n$ lie in a k-dimensional subspace $\mathcal{V}$ with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be written as

$$\mathbf{X} = \mathbf{X}\mathbf{V}\mathbf{V}^T.$$

- $\mathbf{V}\mathbf{V}^T$ is a projection matrix, which projects vectors onto the subspace $\mathcal{V}$.



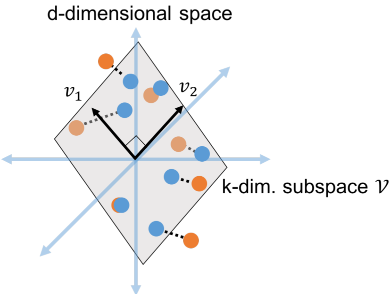
$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Projection View

Claim: If $\vec{x}_1, \ldots, \vec{x}_n$ lie in a k-dimensional subspace $\mathcal{V}$ with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be written as

$$\mathbf{X} = \mathbf{X}\mathbf{V}\mathbf{V}^T.$$

- $\mathbf{V}\mathbf{V}^T$ is a projection matrix, which projects vectors onto the subspace $\mathcal{V}$.

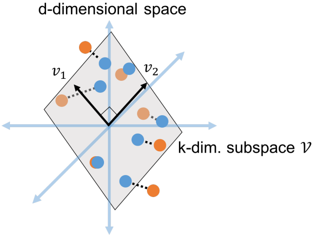


$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Low-Rank Approximation

Claim: If $\vec{x}_1, \ldots, \vec{x}_n$ lie close to a k-dimensional subspace $\mathcal{V}$ with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be approximated as:

$$\mathbf{X} \approx \mathbf{X}\mathbf{V}\mathbf{V}^T$$



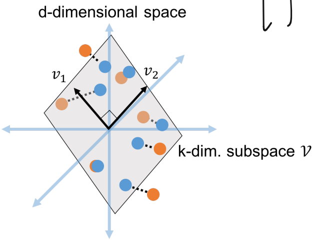
$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Low-Rank Approximation

Claim: If $\vec{x}_1, \ldots, \vec{x}_n$ lie close to a k-dimensional subspace $\mathcal{V}$ with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be approximated as:

$$\mathbf{X} \approx \mathbf{X}\mathbf{V}\mathbf{V}^T$$

$$n\overset{k}{\begin{bmatrix} XV \\ \\ \\ \end{bmatrix}} \overset{d}{\begin{bmatrix} V^T \end{bmatrix}} k$$



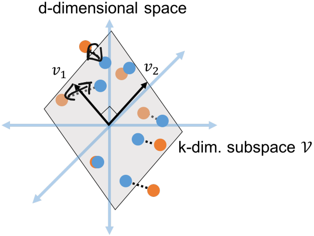
Note: $\mathbf{X}\mathbf{V}\mathbf{V}^T$ has rank k. It is a low-rank approximation of $\mathbf{X}$.

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Low-Rank Approximation

Claim: If $\vec{x}_1, \ldots, \vec{x}_n$ lie close to a k-dimensional subspace $\mathcal{V}$ with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be approximated as:

$$\mathbf{X} \approx \mathbf{X}\mathbf{V}\mathbf{V}^T$$



Note: $\mathbf{X}\mathbf{V}\mathbf{V}^T$ has rank k. It is a low-rank approximation of $\mathbf{X}$.

$$\mathbf{X}\mathbf{V}\mathbf{V}^{\mathsf{T}} = \underset{\mathbf{B} \text{ with rows in } \mathcal{V}}{\arg\min} \|\mathbf{X} - \mathbf{B}\|_F^2 = \sum_{i,j} (\mathbf{X}_{i,j} - \mathbf{B}_{i,j})^2.$$

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Low-Rank Approximation

So Far: If $\vec{x}_1, \ldots, \vec{x}_n$ lie close to a k-dimensional subspace $\mathcal{V}$ with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be approximated as:

$$\mathbf{X} \approx \mathbf{X}\mathbf{V}\mathbf{V}^T.$$

This is the closest approximation to $\mathbf{X}$ with rows in $\mathcal{V}$ (i.e., in the column span of $\mathbf{V}$).

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Low-Rank Approximation

So Far: If $\vec{x}_1, \ldots, \vec{x}_n$ lie close to a k-dimensional subspace $\mathcal{V}$ with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be approximated as:

$$\mathbf{X} \approx \mathbf{X}\mathbf{V}\mathbf{V}^T.$$

This is the closest approximation to $\mathbf{X}$ with rows in $\mathcal{V}$ (i.e., in the column span of $\mathbf{V}$).

- Letting $(\mathbf{X}\mathbf{V}\mathbf{V}^T)_i, (\mathbf{X}\mathbf{V}\mathbf{V}^T)_j$ be the i^{th} and j^{th} projected data points,

$$\|(\mathbf{X}\mathbf{V}\mathbf{V}^T)_i - (\mathbf{X}\mathbf{V}\mathbf{V}^T)_j\|_2 = \|[(\mathbf{X}\mathbf{V})_i - (\mathbf{X}\mathbf{V})_j]\mathbf{V}^T\|_2 = \|[(\mathbf{X}\mathbf{V})_i - (\mathbf{X}\mathbf{V})_j]\|_2.$$

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Low-Rank Approximation

So Far: If $\vec{x}_1, \ldots, \vec{x}_n$ lie close to a k-dimensional subspace $\mathcal{V}$ with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be approximated as:

$$\mathbf{X} \approx \mathbf{X}\mathbf{V}\mathbf{V}^T.$$

This is the closest approximation to $\mathbf{X}$ with rows in $\mathcal{V}$ (i.e., in the column span of $\mathbf{V}$).

- Letting $(\mathbf{X}\mathbf{V}\mathbf{V}^T)_i, (\mathbf{X}\mathbf{V}\mathbf{V}^T)_j$ be the i^{th} and j^{th} projected data points,
$$\|(\mathbf{X}\mathbf{V}\mathbf{V}^T)_i - (\mathbf{X}\mathbf{V}\mathbf{V}^T)_j\|_2 = \|[(\mathbf{X}\mathbf{V})_i - (\mathbf{X}\mathbf{V})_j]\mathbf{V}^T\|_2 = \|[(\mathbf{X}\mathbf{V})_i - (\mathbf{X}\mathbf{V})_j]\|_2.$$
- Can use $\mathbf{X}\mathbf{V} \in \mathbb{R}^{n \times k}$ as a compressed approximate data set.

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Low-Rank Approximation

So Far: If $\vec{x}_1, \ldots, \vec{x}_n$ lie close to a k-dimensional subspace $\mathcal{V}$ with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be approximated as:

$$n \overset{d}{\boxed{\mathbf{X}}} \approx \mathbf{X}\mathbf{V}\mathbf{V}^T. \xrightarrow{} n \overset{k}{\boxed{\mathbf{X}\mathbf{V}}}$$

This is the closest approximation to $\mathbf{X}$ with rows in $\mathcal{V}$ (i.e., in the column span of $\mathbf{V}$).

- Letting $(\mathbf{X}\mathbf{V}\mathbf{V}^T)_i, (\mathbf{X}\mathbf{V}\mathbf{V}^T)_j$ be the i^{th} and j^{th} projected data points,

$$\|(\mathbf{X}\mathbf{V}\mathbf{V}^T)_i - (\mathbf{X}\mathbf{V}\mathbf{V}^T)_j\|_2 = \|[(\mathbf{X}\mathbf{V})_i - (\mathbf{X}\mathbf{V})_j]\mathbf{V}^T\|_2 = \|[(\mathbf{X}\mathbf{V})_i - (\mathbf{X}\mathbf{V})_j]\|_2.$$

- Can use $\mathbf{X}\mathbf{V} \in \mathbb{R}^{n \times k}$ as a compressed approximate data set.

Key question is how to find the subspace $\mathcal{V}$ and correspondingly $\mathbf{V}$.

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Properties of Projection Matrices

$V^T V = I$ $\quad \left[X \right] = \left[\begin{matrix} C \\ c_1 \cdots c_k \end{matrix} \right] \left[\quad V^T \quad \right]_k$

Quick Exercise: Show that $\mathbf{V}\mathbf{V}^T$ is idempotent. I.e., $(\mathbf{V}\mathbf{V}^T)(\mathbf{V}\mathbf{V}^T)\vec{y} = (\mathbf{V}\mathbf{V}^T)\vec{y}$ for any $\vec{y} \in \mathbb{R}^d$.

Why does this make sense intuitively?

Less Quick Exercise: (Pythagorean Theorem) Show that:

$$\|\vec{y}\|_2^2 = \|(\mathbf{V}\mathbf{V}^T)\vec{y}\|_2^2 + \|\vec{y} - (\mathbf{V}\mathbf{V}^T)\vec{y}\|_2^2.$$